빅 데이터 처리를 위한 적응적 사용자 및 토픽 모델링 기반 자동 TV 프로그램 추천시스템

*김은희 **김문철

한국과학기술원

*lins77@kaist.ac.kr **mkim@ee.kaist.ac.kr

Adaptive User and Topic Modeling based Automatic TV Recommender System for Big Data Processing

*Kim, EunHui **Kim, Munchurl

Korea Advanced Institute Science and Technology

요약

최근 TV 서비스의 가입자 및 TV 프로그램 콘텐츠의 급격한 증가에 따라 빅데이터 처리에 적합한 추천 시스템의 필요성이 증가하고 있다. 본 논문은 사용자들의 간접 평가 데이터 기반의 추천 시스템 디자인 시, 누적된 사용자의 과거 이용내역 데이터를 저장하지 않고 새로 생성된 사용자 이용내역 데이터를 학습하는 효율적인 알고리즘이면서, 시간 흐름에 따라 사용자들의 선호도 변화 및 TV 프로그램 스케줄 변화의 추적이 가능한 토픽 모델링 기반의 알고리즘을 제안한다. 빅데이터 처리를 위해서는 분산처리 형태의 알고리즘을 피할 수 없는데, 기존의 연구들 중 토픽 모델링 기반의 추론 알고리즘의 병렬분산처리 과정 중에 핵심이 되는 부분은 많은 데이터를 여러 대의 기계에 나누어 병렬분산 학습하면서 전역변수 데이터를 동기화하는 부분이다. 그런데, 이러한 전역데이터 동기화 기술에 있어, 여러 대의 컴퓨터를 병렬분산처리하기위한 하둡 기반의 시스템 및 서버-클라이언트간의 중재, 고장 감내 시스템 등을 모두 고려한 알고리즘들이 제안되어 왔으나, 네트워크 대역폭 한계로 인해 데이터 증가에 따른 동기화 시간 지연은 피할 수 없는 부분이다. 이에, 본 논문에서는 빅데이터 처리를 위해 사용자들을 클러스터링하고, 클러스터별 제안 알고리즘으로 전역데이터 동기화를 수행한 것과 지역 데이터를 활용하여 추론 연산한 결과, 클러스터별 지역별 TV프로그램 시청 토큰 별 은닉토픽 할당 테이블을 유지할 때 추천 성능이 더욱 향상되어 나오는 결과를 확인하여, 제안된 구조의 추천 시스템 디자인의 효율성과 합리성을 확인할 수 있었다.

1. 서론

최근 Informa Telecom보고에 따르면 전 세계 Connected TV 서비스 이용자는 2016년까지 전 세계 가구의 5억7천만에 이른다는 보고가 있다 [1]. 국내의 IPTV서비스 가입자도 전체 1000만을 넘었으며, 각 서비스 사업자 별로 TV프로그램 콘텐츠의 수도 최대 20만개 가량의 콘텐츠가 제공되고 있다. 이러한 TV서비스 사용자의 증가 및 콘텐츠 수의 증가에 맞춘 빅데이터 처리에 적합한 추천 시스템의 합리적인 디자인이 필요하다.

본 논문은 사용자들의 간접 평가 데이터 기반의 추천 시스템 디자인 시, 누적된 사용자의 과거 이용내역 데이터를 저장하지 않으면서, 새로 생성된 사용자 이용내역 데이터를 학습하는 효율적인 알고리즘이면서, 시간 흐름에 따라 사용자들의 선호도 변화 및 TV 프로그램 스케줄 변화의 추적이 가능한 토픽 모델링 기반의 알고리즘을 제안한다 [3].

본 논문에서는, 빅데이터 처리를 위해서 전체 사용자를 클러스터링 후, 클러스터별로 제안 모델을 활용하여 추천 순위정렬 모델에 필요한 파라미터를 학습하는 구조의 추천 시스템을 통해, 보장된 연산시간과 함께 향상된 추천 성능을 확인할 수 있었다.

2. 관련 연구

사용자의 추천 시스템의 디자인에 있어 전통적으로 이용되어온 추천 알고리즘들은 전체 아이템을 특징 벡터로 두거나, 장르 및 채널 정보와 같은 이미 알려져 있는 고정된 카테고리를 특징 벡터로 주로 사용하여 왔다 [2]. 이미 알려진 장르 및 채널을 기준으로 한 특징 벡터는 사용자들의 실제 구매 욕구 및 시청 취향을 드러내는 데 한계가 있다. 이에 추천 알고리즘으로 토픽모델링을 사용하는 것은 사용자의 실제 구매욕구나 시청취향에 근접한 은닉된 시청 취향을 찾아내는 데 있어 효과가 있다. 또한, 콘텐츠 및 아이템의 증가에 따른 빅데이터 처리에 있어 전체 아이템을 특징 벡터로 사용하는 것에 비해, 효율적인 연산이 가능한 장점을 지닌다.

토픽 모델링 기반의 추론 기술의 빠른 연산을 위해 크게 두 방향의 연구흐름이 있다. 하나는, MCMC (Markov Chain Monte Carlo)를 통한 은닉토픽 추론 중에 할당되는 토픽의 범위가 특정 토픽 범위로 제한된다는 특징을 살려, 일정 연산 후, 토픽 할당 범위를 제한하여 연산을 연속하여 처리속도를 향상 시키는 것이다 [7],[8]. 다른 하나는, 여러 코어로 디자인된 병렬처리 시스템을 활용하여 document (user) 단위의 병렬 추론을 하는 것이다 [4],[5],[6]. 그런데, 이러한 추론 연산 방

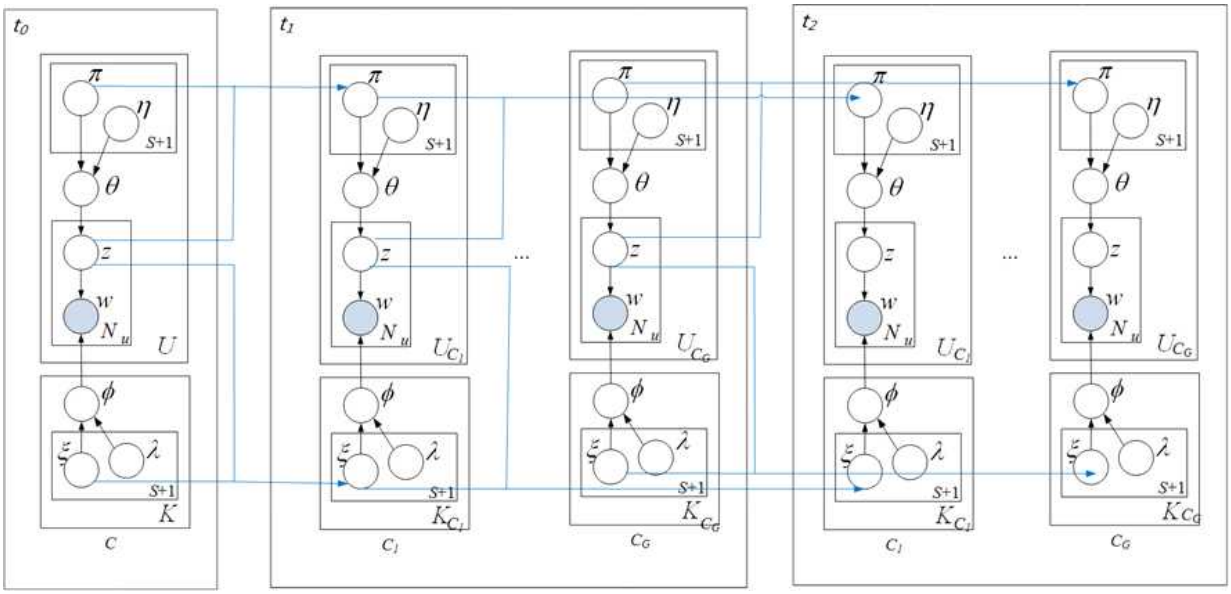

Fig. 1 Adaptive User and Topic Modeling with Clustering

법은 그 처리를 모두 document 단위로 사용한 것에 반해, 제안된 알고리즘은 추론 연산을 추천 시점(epoch)별로 처리하였으며, Sequential Monte Carlo 알고리즘을 사용하여 시간 흐름에 따라 그 연산의 효율성과 효과성을 가져온 장점을 지닌다 [3].

또한, 빅데이터 처리를 위해서는 빠른 추론 연산 외에, 하나의 시스템 가용 처리 범위를 넘어선 데이터에 대해서 이를 시스템별 가용 범위로 나누어 연산하는 분산 컴퓨팅이 필요하다. 이때, 토픽 모델링 기반의 추론 알고리즘을 여러 대의 컴퓨터로 나누어 병렬 분산처리하게 될 경우, 추론 변수 연산을 위해 필요한 전체 단어별 토픽 개수 테이블의 여러 시스템에(머신)에서의 동기화 연산이 필요하다. 이에 Smola 등은 전체 데이터 동기화 과정 중에 하나의 시스템이 모든 데이터들 수집하는 동안 다른 시스템들이 대기하는 방법을 사용하였다[4]. 동기화 과정의 대기 시간 증가를 개선하기 위해, Ahmed 등은 전체 데이터 동기화가 이뤄지지 않더라도 연산 진행이 가능하도록 하였는데(stale synchronization), 전체 단어 데이터를 나누어 여러 대의 서버에 할당하고, 여러 대의 서버들과 클라이언트 간의 동기화 처리를 하되 네트워크 허용 범위 내의 처리를 하도록 디자인 하였다 [5]. 그런데, 이러한 방법도 네트워크 허용범위 내의 연산이 진행되므로, 전역 데이터 동기화 병렬 분산 컴퓨팅은 시간 지연 연산을 초래하게 된다. 이러한 전역 데이터 동기화가 추천시스템 활용에 필요한 파라미터 학습에 필수적인 것인지에 대해 의문을 갖고 접근한 결과, 먼저 사용자들을 클러스터링하고, 클러스터 단위의 추론을 수행하였을 경우, 글로벌 데이터 동기화를 수행한 것에 비해 추천 시스템 정확도가 향상되는 것을 확인할 수 있었다.

3. 클러스터링 기반 적응적 사용자 및 토픽 모델링

가. 제안 모델의 배경 및 도식 모델

시간 흐름에 따라 TV 프로그램 편성표는 변화한다. 특히, 시즌별 TV프로그램 편성의 변화는 많은 TV프로그램 데이터의 소멸과 새로운 프로그램의 등장을 초래한다. 또한, 사용자의 시청 취향 또한 시간 흐름에 따라 변화한다. 그리고, 사용자들은 시청 취향은 시간흐름에 따라 변화한다. 시청률 조사기관 TNmS의 사용자 실제 시청 이용내역 데이터를 기반으로 사용자들의 시청 이용행태를 분석한 결과, 전체 7개월의 시청 기간 중 모든 사용자들이 적어도 1주일 이상은 시청을 멈춘 것을 확인할 수 있었다. 본 논문은 이에 사용자의 선호도 변화 및 TV 프로그램 스케줄 변화를 적응적으로 추적 가능한 알고리즘을 빅데이터 처리를 위해 Fig. 1과 같이 추론된 파라미터 기반 초기 추천 시점에서 사용자별 클러스터링을 수행한다. 이후, 클러스터 단위별 기존 추천 시점의 데이터로 초기화를 수행하는 Sequential Monte Carlo 알고리즘으로 추론의 효율성을 가져오는 알고리즘을 제안한다.

Table 1은 주요 기호에 대한 설명을 나타내는 것이다. 이전 추천 시점에서의 과거 사용자의 시청 기간별 TV프로그램 은닉 토픽에 대한 선호도 및 은닉 토픽 별 TV 프로그램 분포가 계산된 단계에서, 현재 시점(epoch)의 TV시청 과정 모델링은 다음과 같다.

과거 추천 시점 t-1에서의 π_{t-1}, ξ_{t-1} 과 그에 대한 시청 구간 길이별 가중치{λ_t, η_t}를 활용하여,

1. 각 은닉 토픽 k 에 대해,{k=1,2,⋯K},
   (a) k-번째 TV프로그램 시청 취향 그룹의 TV프로그램 선호도 분포를 과거 선호도를 기반으로 이끌어 낸다.

$$,_k \sim \quad ir \sum_{s=0} \lambda_{t,k,s}\xi_{t-1,k,s}\Bigg)$$

2. 각 개별 사용자 u 에 대해,{u=1,2,⋯U},
   (a) u-사용자의 은닉 토픽에 대한 개인 선호도를 과거 개인 선호도를 기반으로 이끌어 낸다.

$$\theta_{t,u} \sim Dir\left(\sum_{s=0}^{S} \eta_{t,u,s}\pi_{t-1,u,s}\right)$$

   (a) 사용자별 각 시청 토큰에 대해, w_j, {j=1,2,⋯W},
      i. 시청 토큰 별 은닉 토픽 라벨을 이끌어 낸다.

$$z_{w\,u} \sim Dir\ \theta_u)= k$$

      ii. 개인 사용자의 은닉토픽에 대한 선호도 $\theta_{t,u}$ 및 은닉 토픽내의 TV프로그램 분포(유사 시청 사용자들의 TV 프로그램 선호 분포) $\phi_{t,k}$를 기준으로 TV 프로그램을 선택한다.

$$w_{t,uk}|z_{w_j u} \sim Dir(\theta_u)= k, \theta_u, \phi_k$$

다. 빅데이터 처리를 위한 제안 모델의 추론

Table 1 Notations

기호	의미
	번째 클러스터
K	토픽의 개수, K_c - n번째 클러스터의 토픽의 개수
U	사용자(시청자)의 수 U_{c_n} - 번째 클러스터의 사용자의 수
t	추천 시점으로 본 논문에서는 일주일 단위로 진행됨.
θ	사용자별 은닉토픽에 대한 선호도로 다항 분포
ϕ	은닉 토픽별 TV 프로그램에 대한 분포(유사 시청 사용자들의 TV 프로그램에 대한 선호도 분포)로 다항 분포
z	각 TV프로그램 토큰 별 토픽 인덱스
w	각 TV 프로그램 인덱스
π	현재 사용자별 토픽에 대한 선호도 분포 θ의 선행 정보(prior)로 디리클레 분포이고, 다중시청 구간 길이 S개+1만큼의 분포의 결합
ζ	현재 대중의 토픽 선호도에 대한 분포 ϕ의 선행정보로 디리클레 분포이고, 다중시청 구간 길이 S개+1의 분포의 결합
η,λ	선호도 분포에 대한 가중치로 η,λ는 각각 π, ζ에 대한 가중치

본 논문에서는 MCMC추정 방법의 하나인 Collapsed Gibbs Sampling (CGS)기반의 추론기법을 Expectation Maximization (EM) 방법으로 업데이트하면서 파라미터들을 학습한다[3]. (E) 단계에서 각 TV program 시청 토큰에 대한 은닉토픽 라벨(z)을 할당하고, (M)단계에서 각각의 선호도의 선행정보의 가중치인 η, λ를 학습한다. (E)단계에서 추론된 은닉 토픽 라벨 분포를 기준으로, 개인사용자의 선호도 θ 및 은닉토픽별 TV 프로그램에 대한 분포 ϕ를 Monte Carlo 로 추론한다. (E)단계에서 각 TV program 시청 토큰에 대한 은닉토픽 라벨(z)을 할당하는 수식은 다음과 같다 [3].

$$P\left(z_j = k \mid W_t, Z_{t-j}, \Pi_{t-1}, I_t, \Xi_t, \Lambda_t\right) = \frac{N_{t,u,k-j} + \sum_s \pi_{uks}^{t-1)} \eta_{tus}}{N_{t,u-j} + \sum_s \eta_{t,u,s}} \cdot \frac{N_{t,k,w_j-j} + \sum_s \xi_{k,w_j-j,s}^{(t-1)} \lambda_{tks}}{N_{t,k-j} + \sum_s \lambda_{t,k,s}} \quad (1)$$

(M) 단계에서 사전 확률 부분을 Maximum likelihood로 추정하여 수식을 푼 반복 연산 형식의 업데이트 수식은 다음과 같다.

$$\eta_{tus} = \eta_{tus}^{old} \times \frac{\sum_w \pi_{uks}^{(t-1)} \left[\Psi\left(N_{tuk} + \sum_{s'} \eta_{tus'}^{old} \pi_{uks'}^{(t-1)}\right) - \Psi\left(\sum_{s'} \eta_{tus'}^{old} \pi_{uks'}^{(t-1)}\right)\right]}{\left[\Psi\left(N_{t,k} + \sum_{s'} \eta_{tus'}^{old}\right) - \Psi\left(\sum_{s'} \eta_{tus'}^{old}\right)\right]} \quad (2)$$

$$\lambda_{tks} = \lambda_{tks}^{old} \times \frac{\sum_w \xi_{kws}^{(t-1)} \left[\Psi\left(N_{tkw} + \sum_{s'} \lambda_{tks'}^{old} \xi_{kws'}^{(t-1)}\right) - \Psi\left(\sum_{s'} \lambda_{tks'}^{old} \xi_{kws'}^{(t-1)}\right)\right]}{\left[\Psi\left(N_{t,k} + \sum_{s'} \lambda_{tks'}^{old}\right) - \Psi\left(\sum_{s'} \lambda_{tks'}^{old}\right)\right]} \quad (3)$$

빅데이터 처리를 위해서는 이러한 연산을 위해 먼저 사용자들별로 클러스터 멤버쉽을 기준으로 사용자 클러스터가 나뉘어서 해당 클러스터별로 Fig. 1과 같이 적응적 모델로 학습한다. 여러대의 컴퓨터를 이용한 분산처리연산 시, EM 학습 과정 중에 주요 추론 파라미터가 되는 $N_{t,u,k}$(사용자 u가 시청한 TV프로그램 시청토큰을 기준으로, 사용자 u에게 은닉토픽 k가 할당된 횟수)는 각 지역머신별로 연산이 가능하나, $N_{t,k,w}$(wTV프로그램에 k 은닉 토픽이 할당된 횟수), 다시 표현하면 전체 단어별 토픽 개수 테이블 (TV프로그램 시청 토큰 별 은닉토픽 할당 테이블로 이를 "TPT 테이블"로 호칭하면)은 전역데이터로써 여러 머신간의 동기화를 필요로 한다. 그런데, 앞서 관련 연구 내용에서 분석한 바와 같이, 전역 동기화에는 네트워크 대역폭 허용 범위 내의 데이터 전달이 가능하여 연산의 지연을 초래하지 않을 수 없다. 이에 전체 클러스터에 대한 전역 동기화를 수행하는 것과 클러스터별 각 머신에서 학습한 경우에 대한 비교 실험을 통해 TV 어플리케이션에서 전역 동기화 필요성 여부를 실험을 통해 확인하였다.

라. 제안 모델 기반의 추천을 위한 순위 정렬 모델

각 u사용자에게 추천하는 프로그램 w의 추천 순위는 식(4)와 같이 개인의 토픽에 대한 선호도와 유사 시청 사용자들의 TV 프로그램에 대한 선호도를 반영하여 결정하게 된다.

$$w) \approx \sum_{u=1}^{K} \sum_{k=1}^{W} \log\left(1 + \left(\theta_{t,u,k} \times \phi_{t,k,w} \times C\right)\right) \quad (4)$$

4. 실험 성능 및 분석

가. 실험 데이터

실험에 사용한 데이터는 시청률 조사기관 TNmS의 데이터로 2011년 1월1일부터 7개월간 2,095명의 사용자들의 실제 TV 시청 이용 내역이다. 7개월 동안 전체 TV 프로그램의 개수는 4,313개이며, 전체 시청 토큰의 개수는 3,713,925 이다. 실제 실험에서는 7개월간 시청이 40개를 넘지 못하는 사용자들을 제외한 1,999명을 실험 대상 데이터로 사용하였다. 학습기간은 가장 짧은 학습 기간(span)을 1주일로 하고, 이보다 긴 다음 학습 기간은 보다 짧은 학습기간에 두배씩 증가하도록 구성하여 1주, 2주, 4주, 8주, 그리고 16주 기간에 걸친 5개의 다중 시청 구간 길이를 학습 기간으로 실험하였다. 시간 흐름에 따라 매 일주일마다 추천 검증을 가정하였고, 추천 검증에 포함시킨 사용자는 일주일 동안 20개 이상의 시청이 있는 사용자들을 대상으로 하였다.

나. 제안 모델의 클러스터링을 통한 성능 향상

Fig.1 에 제안된 것과 같이 사용자별 클러스터링을 수행한 것과, 전체 데이터에 대해 학습한 두 가지 경우의 추천 성능에 대한 비교 검증을 진행하였다. 이 때 두 가지 클러스터링 알고리즘으로 사용하였는데, 하나는 성별, 나이와 같이 사용자들이 가입 시에 기입하는 주어진 정보를 활용하여 클러스터링을 수행하는 데모그래픽 클러스터링과, 개별 사용자들의 은닉토픽 선호도 정보(θ)를 기반으로 클러스터링을 수행하는 K-means 클러스터링 방법이다. 두 방법에 대해 각각 4개의 클러스터링(사용자의 성별과 연령 정보를 기반으로 한 데모그래픽 클러스터링 - DC-AG1에서 DC-AG4과 사용자들의 선호도 정보(θ)를 기반으로 K-means 클러스터링 - KC4-1에서 KC4-4)을 수행하여 비교 실험하였다. Fig.2는 클러스터링을 수행하지 않고 전체 사용자에 대해 제안 모델을 학습시킨 결과 G와 두 가지 클러스터링 방법을 이용하여 추천 성능을 검증한 결과이다. Precision을 기준으로 수식(4)에 제공된 순위 정렬 모델로 정렬된 상위 순위 추천 항목 5개의 정확도를 비교 검증하였다. Fig.2를 통해 확인할 수 있듯이, 클러스터별로 추천한 것이 전체사용자를 기준으로 학습한 G에 비해 대체로 추천 성능이 향상되는 것을 확인할 수 있다.

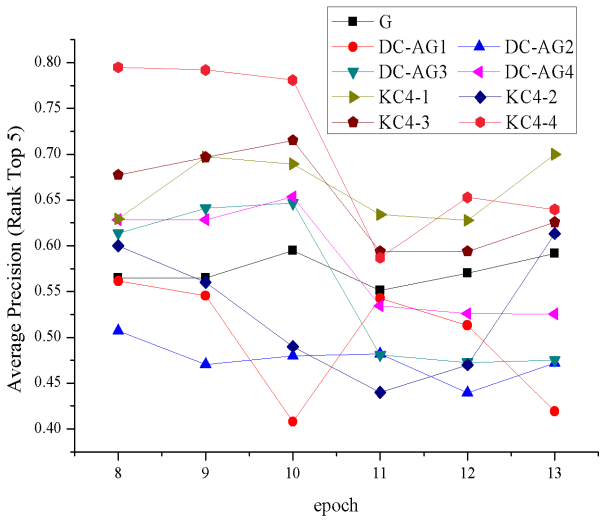

Fig. 2 Precision prediction performance of AUTM for TV user clusters: Rank Top-5 Precision for DC-AG1 - DC-AG4 and KC4-1 - KC4-4

다. 빅데이터 처리의 전역동기화와 지역 동기화 수행

앞서 "3.다"에서 논한 것과 같이 추론 과정 중에 전체 클러스터에 대한 전역 동기화의 필요성 여부를 Fig3.과 같이 확인하였다. 그림을 통해 확인할 수 있듯이, 전역 동기화를 수행한 것(GS1~ GS4)에 비해, 각 클러스터별로 제안모델을 학습한 경우(NS1~NS4)의 추천 성능이 향상되어 나오는 것을 확인할 수 있다. 이를 통해, TV프로그램 어플리케이션 사용을 위한 제안 모델을 활용한 빅데이터 처리에 있어, 전역 동기화 없이 보장된 연산 시간 안에 향상된 추천 성능을 확인할 수 있었다.

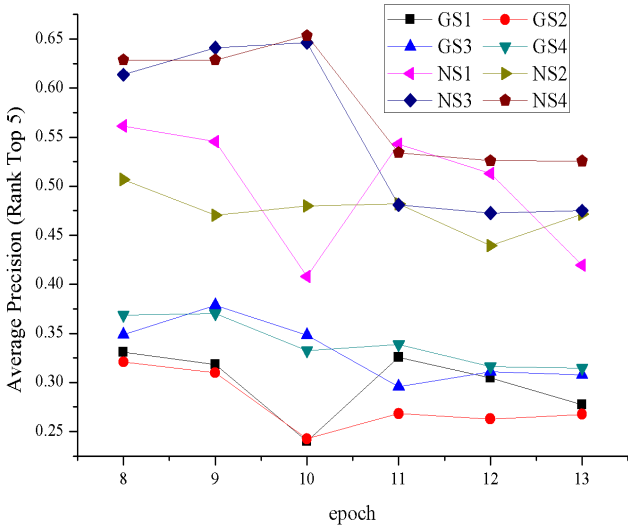

Fig. 3 Precision and recall prediction performance of AUTM for TV user clusters with and without global data synchronization.:Rank Top-5 Precision for *GS*1 - *GS*4 and *NS*1 - *NS*4

5. 결론

본 논문은 실제 산업계 추천 시스템에 사용되는 사용자들의 간접 평가 데이터 기반의 추천시스템으로, 누적된 사용자의 과거 이용내역 데이터를 저장하지 않으면서, 새로 생성된 사용자 이용내역 데이터를 학습하는 효율적인 알고리즘이면서, 시간 흐름에 따라 사용자들의 선호도 변화 및 TV 프로그램 스케줄 변화의 추적이 가능한 토픽 모델링 기반의 알고리즘을 제안하였다. 빅데이터 처리를 위해 사용자들을 클러스터링 하고, 클러스터별 제안 알고리즘으로 전역데이터 동기화를 수행한 것과 지역 데이터를 활용하여 추론 연산한 결과, 클러스터별 'TPT 테이블'을 유지할 때 추천 성능이 더욱 향상되어 나오는 결과를 확인하여, 제안된 구조의 추천 시스템 디자인의 효율성과 합리성을 확인할 수 있었다.

감사의 글

본 논문 연구는 연구재단 중견연구자사업 핵심연구(개인)과제 (과제번호: 2014R1A2A2A01006642)로 수행되었음.

참 고 문 헌

[1] R. Gallagher, G. Cottle, N. Thomson, A. Landbrook et al. "ITM-CES-Connected-TV-White-Paper," Informa Telecoms & MEdia, Feb. ,2012.

[2] G. Adomavicius, and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, June 2005.

[3] E. Kim, S. Pyo and Munchurl Kim, "Adaptive User and Topic Modeling based Automatic TV Recommundation," Summber Conf. Of the Korea Society of Broadcasting Engineers, pp.430-433, Jul. 2012.

[4] A. J. Smola and S. Narayanamurthy, "An Architecture for Parallel Topic Models," Proceedings of the VLDB Endowment, vol. 3, no. 1, pp. 703 - 710, Sept. 2010.

[5] A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. J. Smola. "Scalable inference in latent variable models," Proceedings of the 5th ACM International Conference on Web Search and Data Mining, pp. 123 - 132, Seattle, Washington, USA, Feb. 2012.

[6] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," Journal of Machine Learning Research, vol. 10, pp 1801-1828, 2009.

[7] K. R. Canini, L. Shi, and T. L Griffiths, "Online inference of topics with latent dirichlet allocation," International conference on artificial intelligence and statistics, pp. 65-72, 2009.

[8] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M. "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation," Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 569-577, 2008.

[9] TNmS Research, http://home.tnms.tv/company/main.asp.